HW #5 (Due: May 08, 2023)

Problem 1.

Consider a system with $\gamma = 0.9$ the following state and action spaces: $\mathcal{S} = \{-1, 1, 2\}$, $\mathcal{A} = \{-1, 0, 1\}$. The available batch data are as follows:

$$D = \{ (s_0 = 1,\ a_0 = 1,\ r_1 = 1,\ s_1 = 2),\ (s_1 = 2,\ a_1 = 0,\ r_2 = -1,\ s_2 = 1)$$
$$(s_2 = 1,\ a_2 = -1,\ r_3 = 0,\ s_3 = -1) \}$$

Consider the basis function $\Phi(s,a) = a^2 s + as + a$ with initial weights $w^0 = 1$. Perform LSPI to compute w^1 and w^2, and policy associated to w^2.

* In case of tie for action selection, give the preference to -1, then 0 and finally +1.

$$\text{Example} \longrightarrow \underset{a \in \{-1, 0, 1\}}{\arg\max} \left\{ \overset{-1}{2} \quad \overset{0}{2} \quad \overset{+1}{2} \right\} = -1$$

Problem 2.

Repeat Problem 1 using the basis function $\Phi(s,a) = \begin{bmatrix} as + a \\ a^2 s \end{bmatrix}$ with initial weights $w^0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$. Perform LSPI to compute w^1 and w^2, and policy associated to w^2. Is the final policy (i.e., π^2) different from Problem 1?

Can two basis functions, in general, lead to different policies?

Appendix: LSPI.

Policy Evaluation

$$\underset{k\times1}{w^-} \longrightarrow Q^{\pi}(s,a) = \Phi^T(s,a)\, w^-$$

$$\pi^-(s) = \underset{a\in A}{\operatorname{argmax}}\, Q^{\pi^-}(s,a) = \underset{a\in A}{\operatorname{argmax}}\, \Phi^T(s,a)\, w^-$$

$w^- = w^+$

Policy Improvement

$$\underset{k\times k}{A} = \frac{1}{L}\sum_{i=1}^{L} \underset{k\times1}{\Phi(s_i,a_i)}\left[\underset{k\times1}{\Phi(s_i,a_i)} - \gamma\, \underset{k\times1}{\Phi(s_{i+1},\pi^-(s_{i+1}))}\right]^T$$

$$\underset{k\times1}{b} = \frac{1}{L}\sum_{i=1}^{L} \underset{k\times1}{\Phi(s_i,a_i)}\, r_{i+1}$$

$$\underset{k\times1}{w^+} = \underset{k\times1}{A^{-1}}\, \underset{k\times1}{b}$$